Narish Vij NPL

A Genetic Algorithm to Select Variables in Logistic Regression: Example in the Domain of Myocardial Infarction

Staal Vinterbo*, M.Sc., Lucila Ohno-Machado, M.D., Ph.D.,
Decision Systems Group and Division of Health Sciences and Technology
Harvard Medical School/Massachusetts Institute of Technology
Boston, Massachusetts

Actual use of regression models in clinical practice depends on model simplicity. Reducing the number of variables in a model contributes to this goal. The quality of a particular selection of variables for a logistic regression model can be defined in terms of the number of variables selected and the model's discriminatory performance, as measured by the area under the ROC curve. A genetic algorithm was applied to search for the best variable combinations for modeling presence of myocardial infarction in a data set of patients with chest pain. Using an external validation set, the resulting model was compared with models constructed with standard backward, forward and stepwise methods of variable selection. The improvement in discriminatory ability yielded by the genetic algorithm variable selection method was statistically significant ($p < 0.02$).

INTRODUCTION

Logistic regression models are common in the field of medicine. Several studies on diagnosis of coronary disease involving logistic regression models have been published [1, 2, 3]. Some of these models were built to be used prospectively on previously unseen cases. These are considered predictive models. Predictive models can be compared in terms of performance, robustness, explanatory power, and cost. Performance is often measured by discriminatory ability (e.g., area under the Receiver Operating Characteristic, or "ROC", curve) and calibration (e.g., plots of expected versus observed results). Robustness can be interpreted as the ability to generalize the model to other data and to maintain good performance in presence of uncertainty and/or missing data items. Explanatory power is the ability to explain certain dependencies in the data and the model results. Cost can be measured as an aggregate of associated costs of obtaining the information.

A factor that contributes to performance, robustness, explanatory power and cost is parsimony of the model. A smaller model (in terms of the number of variables) is likely to (*i*) avoid over-fitting problems, thus performing and generalizing better, (*ii*) be less likely to fail due to missing data, (*iii*) be easier to explain, and (*iv*) cost less, both in terms of data collection and computational effort.

Traditionally, for logistic regression models, this issue has been addressed by stepwise forward, backward, and composite variable selection methods [4]. The SAS statistical software system [5] calls these selection methods "forward", "backward" and "stepwise", respectively. These terms will be used in the remainder of this article. Although being well understood and relatively easy to compute, these methods consider the addition or removal of one variable at the time, conditional on the variables already selected. This sequential approach restricts the examined number of models severely. Another approach is to examine all possible models. Given u variables to choose from, the number of possible models is 2^u, which renders this exhaustive approach infeasible with other than small numbers of variables. The SAS system offers this possibility, but only for 10 or fewer variables.

Heuristic approaches based on genetic algorithms have been used for selection of input variables and other parameters in artificial neural networks [6, 7, 8]. A search through bibliographic databases such as INSPEC, MEDLINE, MathSciNet, Science Citation Index, HealthStar, and Applied Science and Technology Index, together with a multiple web search engine search, did not

*Corresponding author. On leave from the Knowledge Systems Group, Dep. of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway.

reveal any publications that deal with genetic algorithm variable selection for logistic regression. We have implemented a genetic algorithm based variable selection method for logistic regression models, and compare it to traditional sequential variable selection methods using data sets of patients with chest pain. In the next section, we review the basic ideas behind genetic algorithms and explain how we applied them to variable selection.

METHODS

A genetic algorithm is a heuristic for function optimization where the extrema of the function (i.e., minima or maxima) cannot be established analytically. A population of potential solutions is refined iteratively by employing a strategy inspired by Darwinistic evolution or natural selection. Genetic algorithms promote "survival of the fittest".

Given an initial population, often created randomly, the principal steps of a genetic algorithm are:

1. Select parents from the current population to undergo genetic operations to form offspring. This is done stochastically with preference assigned to individuals that yield higher function values (i.e., the "fittest" individuals).

2. Apply genetic operations such as crossover, mutation and inversion to the selected parents to form offspring. The operators are designed such that properties of the parents are reproduced in the offspring.

3. Recombine the offspring and current population to form a new population.

These steps are performed until some predefined stopping criterion is met. The selection method from a population of potential solutions, with preference to "fittest" individuals, has given these types of algorithms the name "genetic", or sometimes "evolutionary", algorithms. The individuals in a population are often called "chromosomes", built out of "genes" that represent the properties of the individual, and the function to optimize is referred to as a "fitness" function. Each iteration is called a "generation". A cycle of this process is shown in Figure 1. A pseudo-code skeleton for a genetic algorithm applying crossover, mutation

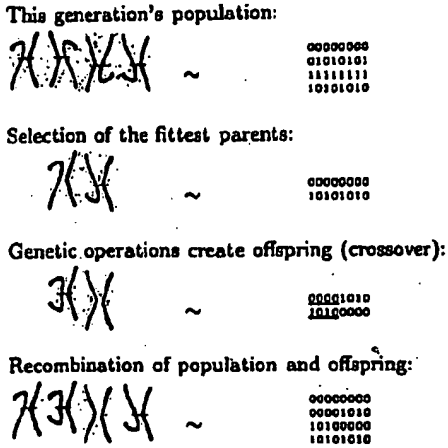

Figure 1: One cycle of evolution. Presented at the left of the figure is a population of four manipulated images of the human chromosome 1. Presented at the right is a bit-vector representation of the "genes", representing the properties of the image manipulations. The two outer chromosomes are selected to be parents, and crossover is applied (after the 4th bit). The two middle chromosomes are then replaced by the offspring to form the next generation population.

and inversion as genetic operators, is shown in Figure 2. For an in depth explanation and discussion of genetic algorithms, see [9, 10].

```
P ← initializePopulation()
evaluate(P)
while(not stop(P)) do
   Parents[1..3] ← selectParents(P)
   Offspring[1] ← Crossover(Parents[1])
   Offspring[2] ← Mutation(Parents[2])
   Offspring[3] ← Inversion(Parents[3])
   P ← recombine(P, Offspring[1..3],
                 Parents[1..3])
   evaluate(P)
done
```

Figure 2: Pseudo-code for the genetic algorithm. `P` denotes the population, `Parents[i]` denotes a set of selected individuals to undergo a genetic operation and `Offspring[i]` denotes the resulting set of individuals.

The objective of variable selection for logistic regression models is to find parsimonious models that perform as well or better than the model that utilizes all available information. With this objective in mind, we construct a measure of fitness for a selection of variables v. Given two tagged sets of data, a training or construction set C, and a hold-

out or selection set S, a logistic regression model $m_C(v)$ can be constructed using C, and evaluated using S. The result is a numeric value $c_S(m_C(v))$ representing the performance of $m_C(v)$ on S. If the total number of variables for is u, and the number of variables in the selection v is n, we propose the following fitness function:

$$(1) \qquad f(v, C, S) = c_S(m_C(v)) + \rho \frac{u - n}{u}.$$

The first term rewards models with good performance, and the second term rewards parsimonious models. The parameter ρ determines the weight that is placed on such a reward.

The genetic algorithm is configured by parameters such as: the fraction of the population to undergo each genetic operation, the size of the population, the fitness function, and the stopping criteria. A predefined number of the best encountered individuals is returned as the result of one run of the algorithm.

EXPERIMENTS

The objective of our experiment was to compare the performance of a logistic regression model constructed using the variable selection method based on the genetic algorithm with models constructed using standard forward, backward and stepwise variable selection.

The models were constructed using a data set from Sheffield, England, of 500 patients with chest pain presenting at the emergency room (ER). The data set contained 43 predictor variables and one outcome, indicating whether these patients had a myocardial infarction (MI) or not. The prevalence of MI was 30%.

For the application of the genetic algorithm, the set was randomly split into a training part C, and a hold-out part S. The parts had 335 and 165 cases, respectively. The chromosomes were represented as binary vectors, where the presence of a bit indicates the presence of the corresponding variable in the logistic model. The "genetic" operators crossover, mutation and inversion were used, and selection was done by universal stochastic sampling. This was also used in the selection of individuals to replace in the fixed size population in the recombination step. Initialization was random, and the stopping criteria was lack of improvement in the average fitness of the population over 20 generations. The population size was set to 70, the probabilities for selection for crossover, mutation and inversion were 0.3, 0.1 and 0.1, respectively. Each "individual" (i.e., combination of variables selected) was transformed into a logistic regression model $m_C(v)$ using the SAS system LOGISTIC procedure. The coefficients were calculated using the training set C. The performance measure $c_S(m_C(v))$ was the area underneath the receiver operating characteristic (ROC) curve [11], computed as its equivalent statistic, the c-index [12] on the hold-out set S. A ρ value of 0.05 was empirically chosen for the fitness function.

The genetic algorithm ran for 79 generations, requiring 1549 fitness function evaluations. The fittest model was selected as the result of the method, and labeled model "g".

The logistic regression models with sequential variable selection were constructed using the SAS system LOGISTIC procedure on the entire set with significance levels for entry and removal of 0.05. They are termed model "f", for forward selection, model "b", for backward selection, and model "s", for stepwise selection. Additionally, a model "a" was constructed with all 43 available variables.

RESULTS

An overview of the variables selected by the different models can be seen in Table 1. The variables "gender", "right arm pain", "diaphoresis", "previous angina", and "ST elevation" were selected by all methods. Certain variables that were consistently selected by the sequential methods, such as "sharp pain", "episodic pain", "hypoperfusion", and "ST or T abnormality" were not selected by the genetic algorithm method.

The final models were evaluated on an external validation set of 1253 cases collected in Edinburgh, Scotland. The resulting c-indices were statistically compared using the method of Hanley and McNeil [13]. The results are in Table 2. Our model "g" was significantly better ($p < 0.02$) than any of the other models evaluated on the external validation set. There were no statistically significant differences between the other models. The corresponding ROC curves for all the models are shown in Figure 3.

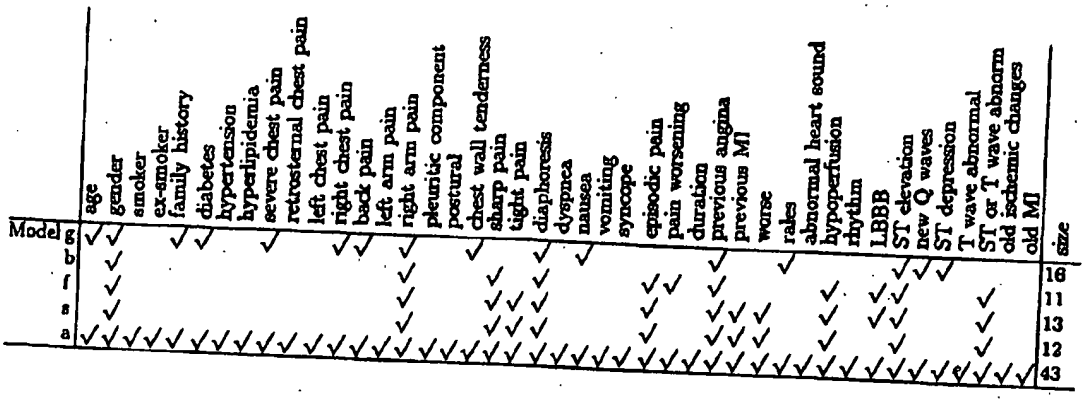

Table 1: The input variables selected by the different methods. Model name "g" indicates the genetic algorithm choice of input variables, while "b", "f", "s", and "a" indicate backward, forward, stepwise selection, and all variables included, respectively.

model	c-index	se	H-M p
g	0.939	0.0086	N/A
b	0.916	0.0103	0.0178
f	0.910	0.0108	0.0074
s	0.907	0.0110	0.0038
a	0.916	0.0099	0.0018

Table 2: For all models: c-index, standard error (se) and test p value (H-M p) for difference in c-indices between a given model and model "g". These results were obtained by evaluating the models on an independent validation set.

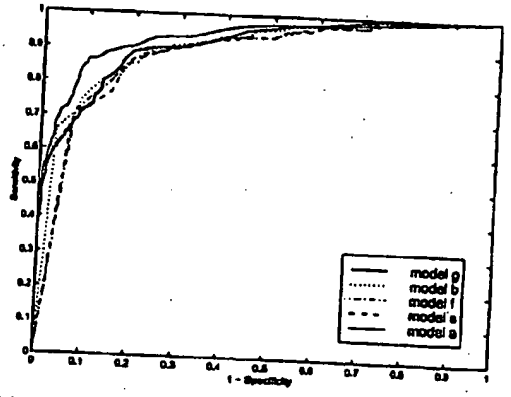

Figure 3: ROC curves for the different logistic models on an independent evaluation set.

DISCUSSION

Although the computational effort spent by the genetic algorithm evaluating the 1549 different variable combinations is considerably larger than the effort spent on doing sequential variable selection, it is still on the order of the effort spent by the SAS system when using the "best" selection option for just 10 variables. The SAS system will, with the "best" selection option, do an exhaustive search through all 1024 possible variable combinations from a set of 10 variables. The number of models that an exhaustive search of all possible variable combinations from 43 variables requires is 2^{43} (on the order of 10^{13}), a number much larger than 1549.

The sequential selection methods agreed to a high degree on the variables to include. The genetic-algorithm-based method selected some variables selected by the sequential methods, but also others. Curiously it selected variables such as "diabetes", "severe chest pain", "ST elevation", "new Q waves", and "ST depression", also deemed important by cardiologists [14].

A finding that supports the importance of parsimony of models, outlined in the introduction, was that the selective pressure on smaller models (second term in equation 1) seemed to be necessary for the success of the method. Using only the c-index (area under the ROC curve) as the fitness function did not produce a selection that was significantly better than any of the sequential variable selection methods.

In order to generalize these results it is necessary to validate the genetic-algorithm-based method rigorously using techniques such as cross-validation and/or bootstrapping. However, the use of an independent external validation set de-

creases the chances that the model is over-fitting the data, and strongly suggests that generalization to other data is warranted. We plan to test the presented method in other domains.

Another future area of investigation is the effect of changing parameters such as ρ in the genetic algorithm, and comparing the resulting models with models created with sequential selection using a wider range of entry/removal levels.

CONCLUSION

We have presented a genetic-algorithm-based variable selection method for a logistic regression that models the presence of myocardial infarction in a patient presenting at the ER with chest pain. The improvement of discriminatory performance achieved by this method was statistically significant ($p < 0.02$) over models constructed using traditional variable selection methods.

Acknowledgments

This work was supported in part by project grant 107409/320 from the Norwegian Research Council, by contract 467-MZ-802289 from NLM/NHLBI, and by grant R29 LM06538-01 from the National Library of Medicine.

We wish to thank Hamish Fraser for providing the data sets.

References

[1] H. P. Selker, J. R. Beshansky, J. L. Griffith, T. P. Aufderheide, et al. Use of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac ischemia. A multicenter, controlled clinical trial. *Annals of Internal Medicine*, 129(11):845–55, 1998.

[2] D. Do, J. A. West, A. Morise, E. Atwood, and V. Froelicher. A consensus approach to diagnosing coronary artery disease based on clinical and exercise test data. *Chest*, 111(6):1742–9, 1997.

[3] R. L. Kennedy, A. M. Burton, H. S. Fraser, L. N. McStay, and R. F. Harrison. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: Derivation and evaluation of logistic regression models. *European Heart Journal*, 17:1181–1191, 1996.

[4] R. Christensen. *Log-Linear Models and Logistic Regression*. Springer Verlag, New York, 2nd edition, 1997.

[5] SAS Institute Inc., Cary, NC, USA. *SAS/STAT User's Guide, Version 6*, 4th edition, 1989.

[6] M. N. Narayanan and S. B. Lucas. A genetic algorithm to improve a neural network to predict a patient's response to warfarin. *Methods of Information in Medicine*, 32(1):55–8, 1993.

[7] R. Dybowski, P. Weller, R. Chang, and V. Gant. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *Lancet*, 347(9009):1146–50, 1996.

[8] M. F. Jefferson, N. Pendleton, S. B. Lucas SB, and M. A. Horan. Comparison of a genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. *Cancer*, 79(7):1338–42, 1997.

[9] Z. Michalewicz. *Genetic Algorithms + data structures = evolution programs*. Springer Verlag, New York, 1992.

[10] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT press, Cambridge, 1996.

[11] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[12] F. E. Harrell Jr., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982.

[13] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843, 1983.

[14] S. Dreiseitl, L. Ohno-Machado, and S. Vinterbo. Evaluating variable selection methods for diagnosis of myocardial infarction. Submitted to the 1999 AMIA annual symposium.